A. Supplementary Material

A.1. Proof of Lemma 1

Proof. Consider the αvector induced when agents follow plan $a_{t:\ell-1}$ from control interval t onwards, denoted $\alpha^{a_{t:\ell-1}}$ and given by $\alpha^{a_{t:\ell-1}}(x,o) \doteq \mathbb{E}\{R_t | x_t = x, o_t = o, a_{t:\ell-1}\}$. Hence, the optimal value starting at any occupancy state s_t is given by taking the maximum over values of all possible plans from control interval t onwards: $V_t^*(s_t) = \max_{a_{t:\ell-1}} \langle s_t, \alpha^{a_{t:\ell-1}} \rangle$. In addition, the linearity of the expectation also implies that $\alpha^{a_{t:\ell-1}}$ is linear in the occupancy-state space. The proof directly follows from (Rockafellar, 1970, Theorem 5.5). □

A.2. Proof of Lemma 2

Proof. We proceed by induction to prove this property. In the following we assume that all operations (e.g. integrals) are well-defined in the corresponding spaces. For control interval $t = \ell - 1$, we only have to take into account the immediate reward and, thus, we have that $Q^*_{\ell-1}(s_{\ell-1}, a_{\ell-1}) = R(s_{\ell-1}, a_{\ell-1})$. Therefore, if we define the set $\Omega^*_{\ell-1} = \{q_{\ell-1}\}$, where $q_{\ell-1}(x,o,u) \doteq r(x,u)$, the property holds at control interval $t = \ell - 1$. We now assume the property holds for control interval $\tau + 1$ and we show that it also holds for control interval τ. Using (2) and (4), we have that, $Q^*_\tau(s_\tau, a_\tau) = R(s_\tau, a_\tau) + \gamma_1 \max_{a_{\tau+1}} Q^*_{\tau+1}(T(s_\tau, a_\tau), a_{\tau+1})$, and by the induction hypothesis, let $s_{\tau+1} \doteq T(s_\tau, a_\tau)$:

$$Q^*_{\tau+1}(s_{\tau+1}, a_{\tau+1}) = \max_{q \in \Omega^*_{\tau+1}} \sum_{x,o,u} s_\tau(x,o) a_\tau(u|o) \sum_{y,z,u'} p^{u,z}(x,y) a_{\tau+1}(u'|o,u,z) q(y,(o,u,z),u').$$

With the above,

$$Q^*_\tau(s_\tau, a_\tau) = \max_{a \in A_{\tau+1}, q \in \Omega^*_{\tau+1}} \sum_{x,o,u} s_\tau(x,o) a_\tau(u|o)[r(x,u) + \gamma_1 \sum_{y,z,u'} p^{u,z}(x,y) a(u'|o,u,z) q(y,(o,u,z),u')].$$

At this point, we can define the bracketed quantity as

$$q^{a_{\tau+1}}(x,o,u) \doteq r(x,u) + \gamma_1 \sum_{y,z,u'} p^{u,z}(x,y) a_{\tau+1}(u'|o,u,z) q(y,(o,u,z),u').$$

Note that α-vector $q^{a_{\tau+1}}$ is independent of occupancy state s_τ and decision rule a_τ for which we are computing Q^*_τ. With this, we have that $Q^*_\tau(s_\tau, a_\tau) = \max_{q^a \colon a \in A_{\tau+1}, q \in \Omega^*_{\tau+1}} \langle s_\tau \odot a_\tau, q^a \rangle$ and, thus the lemma holds. □

A.3. Proof of Theorem 1

Proof. The proof derives directly from Lemma 2. First, notice that any arbitrary non-dominated joint plan ρ induces a sequence of α-vectors $q^\rho_{0:\ell-1}$ stored in $\Omega^*_{0:\ell-1}$, which proves the Q-value function under a fixed plan is linear over occupancy states and joint decision rules. In addition, each α-vector $q^\rho_t \in \Omega^*_t$ describes the expected returns from $t \in [\![0; \ell-1]\!]$ onward, when agents follow non-dominated joint plan ρ. If we let ρ^* be a greedy joint plan with respect to $Q^*_{0:\ell-1}$, then $q^{\rho^*}_{0:\ell-1}$ is maximal along $\{T(s_0, a^*_{0:t-1})\}_{t \in [\![0;\ell-1]\!]}$. □